认知诊断评价中的被试拟合研究

喻晓锋[**1] 唐茜[1] 秦春影[2] 李喻骏[1]
（[1]江西师范大学心理学院，南昌，330022）
（[2]南昌师范学院数学与信息科学学院，南昌，330032）

摘 要 通常情况下，认知诊断需要通过认知诊断模型对被试进行诊断评价。认知诊断模型所生成的诊断结果的有效性依赖于被试作答反应是否与所选用的模型拟合。因此，在对诊断结果进行评估的时候，需要通过被试拟合分析来对被试个体的作答反应与模型的拟合情况进行检验，以避免错误或无效的补救措施。本研究基于加权的得分残差，提出认知诊断评价中新的被试拟合指标R。模拟研究结果表明，R指标的一类错误率有较好的稳定性，对随机作答、疲劳、睡眠和创造性作答四种异常被试类型均有较高的统计检验力。并将R指标应用于分数减法实证数据，展示R指标在实际测验中的使用过程。

关键词 认知诊断 被试拟合 DINA 模型 异常反应作答

1. 引言

近年来，认知诊断评价（cognitive diagnosis assessment, CDA）在心理和教育测量中得到了广泛的应用，它对考生是否掌握知识点或技能进行分析，为进一步学习和教学提供指导（Leighton & Gierl, 2007; Rupp et al., 2010; Tatsuoka, 1983）。认知诊断模型在认知诊断评价过程中起到统计工具的作用，它被用来推断被试所掌握的属性，诊断过程需要借助它来对被试进行诊断评价（von Davier & Lee, 2019）。认知诊断模型与测验数据的拟合情况，直接影响依据这个模型得到的诊断结果的准确性，并影响整个测验的信度和效度，因此认知诊断评价需检验模型-资料拟合优度。标准 5.19 （the *Standards for educational and psychological testing,* pp107）中明确提出在教育和心理测量中，需要对所选择的项目反应模型与作答反应数据进行拟合检验。

在教育测量中，考试分数是用来衡量被试的能力水平的，但由于被试可能存在的异常行为，考试分数不一定是被试技能或知识的真实反映。在心理测量学中，衡量被试的实际作答反应与其模型预测的反应之间的差异的方法称为被试拟合（Meijer & Sijtsma, 2001）。被试拟合用来检验被试个人作答反应与认知诊断模型的拟合程度，恰当的认知诊断模型应该比较准确的反映被试在项目反应过程中的心理加工特征，以有效地推断被试属性掌握情况。被试作答反应能够拟合所选择的认知诊断模型，称为被试拟合（person-fit）；反之，如果被试出

*本研究得到江西省教育科学十四五规划 2021 课题（21YB257）的资助
**通讯作者：喻晓锋，E-mail：xyu6@jxnu.edu.cn；

现异常作答反应，和所选择的认知诊断模型不拟合，称为被试不拟合（person-misfit）。如果出现被试不拟合，一方面，根据失拟被试的作答反应数据对其属性掌握模式进行推断的结果可能是难以解释或无效的，进一步导致不合适的补救措施，其次，失拟被试的数据可能会影响整个测验的信效度，因此被试拟合检验尤为重要。以往关于被试拟合的研究大多集中在项目反应理论（item response theory, IRT; Baker & Kim, 2004）下开展，在认知诊断评价中，被试拟合检验在测验评价分析过程中较易被忽视，与被试拟合有关的研究较少。目前已有的研究主要包括：Liu 等人（2009）基于边际和联合似然比检验，提出了用于判别异常作答被试的似然比检验统计量，引入异常反应概率变量 ρ_i，并用标示变量A定义异常反应被试类型，其局限性在于实践过程中异常被试和异常反应类型较难被人为定义；Cui 和 Leighton（2009）开发了在属性层级模型下衡量被试观察反应模式和理想反应模式是否匹配的层级一致性指标（hierarchical consistency index, HCI），层级一致性指标基于属性层级模型，即强调属性间的关系，当测验所考察的属性之间只有部分属性具有层级关系或者属性之间没有层级关系时，HCI指标就不适用；Liu 等人（2009）提出的似然比检验统计量被证明在使用 DINA 模型时对虚假的高分（spuriously high scores）和虚假的低分（spuriously low scores）具有较好的检测力；Cui 和 Li（2015）将l_z指标扩展到认知诊断框架下，同时提出了一种新的比较观察反应模式和理想反应模式的反应一致性指标（response conformity index, RCI）；还有研究者对认知诊断测验中的被试拟合检验进行了综述和分析（陈孚等, 2016; 涂冬波等, 2014）。正是因为诊断测验中被试拟合研究的重要性，本研究拟构建基于认知诊断测验的被试拟合指标，并将它与l_z和RCI指标进行比较，考察它们在不同条件下的表现。有关$\mathrm{l_z}$和RCI指标的介绍，请见附录 A。

2. 认知诊断评价下被试拟合指标R的提出

残差是回归分析中的重要概念，残差在数理统计中是指实际观察值与期望值（拟合值）之间的偏差。残差应用其中蕴含的逻辑就是，通过对比理想情况与实际情况的差异而发现其中的异常情况。预期偏差会使残差统计量膨胀，这与被试拟合检验的思想一致。本研究打算构建基于残差的被试拟合统计量R指标来进行诊断测验中的被试拟合分析。下面首先给出标准化残差的定义。

2.1 标准化残差的定义

在 IRT 有关的很多研究中，尤其是有关 Rasch 模型的研究，有很多和标准化残差$\frac{x_{ij}-E(X_{ij}|\theta_i)}{\sqrt{Var(X_{ij}|\theta_i)}}$有关的应用（Masters & Wright, 1997）。其中$Var(X_{ij}|\theta_i)$是给定能力值θ_i随机变量

X_{ij}的方差。对考生在各项目上标准化残差求和之后就可以作为被试拟合的评价指标。一方面，标准化残差可以看作是一种加权的残差，权重是项目作答的条件标准误的倒数，它近似服从标准正态分布。另一方面，因为被试拟合关注的是考生的观察作答与模型的预测作答之间的一致性，当观察作答与模型的预测之间存在严重的不一致时，表现在出现这个观察作答的概率很小，并且由于它处于分母的位置，是一个逆向的权重，就会导致残差的取值虚高，因此基于以上的考虑，本研究以观察作答概率的倒数作为被试拟合统计量的权重，定义新的指标$\boldsymbol{R}$。

2.2 R指标的定义

R指标的数学表达式如下：

$$R_i = \sum_{j=1}^{J} log\left[\frac{x_{ij}-E(X_{ij}|\alpha_i)}{P(x_{ij}|\alpha_i)}\right]^2 \qquad (1)$$

其中，x_{ij}表示被试i在项目j上的观察得分，α_i是被试i的属性掌握模式。在实际应用中，真实的被试属性掌握模式是无法得到的，因此本研究采用被试属性掌握模式估计值。$E(X_{ij}|\alpha_i)$表示属性掌握模式为α_i的被试i在项目j上的期望得分，如在 DINA 模型（de la Torre, 2009）中，每个项目只包含两个参数：失误参数s（slipping parameter）和猜测参数g（guessing parameter）。如果被试i掌握了项目j考察的所有属性，此时$E(X_{ij}|\alpha_i) = 1 - s_j$，如果被试$i$至少有一个项目$j$考察的属性未掌握，此时$E(X_{ij}|\alpha_i) = g_j$，分子是观察作答与期望得分之差。分母$P(x_{ij}|\alpha_i)$表示属性掌握模式为$\alpha_i$的被试$i$在项目$j$上得$x_{ij}$分的概率，当属性掌握模式为$\alpha_i$的被试$i$掌握了项目$j$考察的属性并正确作答时，$P(x_{ij} = 1|\alpha_i) = E(X_{ij}|\alpha_i)$。当$P(x_{ij}|\alpha_i)$值越小时，被试失拟程度越高，它进一步放大了观察作答和期望作答之间的残差。R_i是被试i在所有项目上的R值的和，其值越大表示越不拟合；而对于一个“拟合良好”的被试来说，可以预期其R_i值相对更小。需要注意的是，R指标本身不依赖于特定的诊断模型，因为 DINA 模型具有参数简单、易于使用、有很多的开源软件都包含 DINA 模型，这是选择 DINA 模型作为实例的原因，有关 DINA 模型的具体信息请参考（de la Torre, 2009; Junker & Sijtsma, 2001; von Davier & Lee, 2019）。

在 DINA 模型中，则对于每个考生来说，他/她所完成的项目根据其对属性的掌握情况和作答情况可以分成四类：掌握某项目考察的属性，但是正确作答（η_{11}）或错误作答（η_{10}）；未完全掌握某项目，错误作答（η_{00}）或正确作答（η_{01}）。这里η表示对应类型的题目数量，它的第一个下标表示被试对项目属性的是否完全掌握，第二个下标表示其作答是否正确，它的取值为 1 表示完全掌握或正确作答。这样一来，公式 1 可以写成如下的形式：

$$R_i = \sum_{j=1}^{J_{11}} log\left[\frac{s_j}{1-s_j}\right]^2 + \sum_{j=1}^{J_{10}} log\left[\frac{1-s_j}{s_j}\right]^2 + \sum_{j=1}^{J_{00}} log\left[\frac{g_j}{1-g_j}\right]^2 + \sum_{j=1}^{J_{01}} log\left[\frac{1-g_j}{g_j}\right]^2, \quad (2)$$

其中，J_{11}，J_{10}，J_{00}和J_{01}分别对应η_{11}，η_{10}，η_{00}和η_{01}的题目数。进一步，当s_j和g_j都小于 0.5 时，公式 2 可以变换成：

$$R_i = 2\left\{\sum_{j=1}^{J_{11}} \log\left[\frac{s_j}{1-s_j}\right] + \sum_{j=1}^{J_{10}} \log\left[\frac{1-s_j}{s_j}\right] + \sum_{j=1}^{J_{00}} log\left[\frac{g_j}{1-g_j}\right] + \sum_{j=1}^{J_{01}} log\left[\frac{1-g_j}{g_j}\right]\right\}, \quad (3)$$

可以看出，对于一个“拟合良好”的被试来说，J_{10}和J_{01}应该会小，$\log\left[\frac{s_j}{1-s_j}\right]$和$log\left[\frac{g_j}{1-g_j}\right]$都是负值，此时$R_i$的值应该是较小的值；对于“拟合不好”的被试，$J_{10}$和$J_{01}$的值相对会更大，$\log\left[\frac{1-s_j}{s_j}\right]$和$log\left[\frac{1-g_j}{g_j}\right]$都是正值，此时$R_i$的值应该是更大的值。

3. 研究一：$\boldsymbol{R}$指标与l_z、RCI指标比较研究

为了评价R指标在诊断测验的被试拟合检验上的表现，我们展开模拟研究来比较R指标与l_z指标、RCI指标的表现。Cui 和 Li（2015）的研究表明，RCI指标优于 Liu 等（2009）的似然比统计量，故没有将似然比统计量作为比较对象。

3.1 研究设计

探讨在 DINA 模型下，不同项目长度、项目质量、失拟被试类型下R_i指标和l_z、RCI指标的一类错误率和统计检验力。项目长度和质量是影响诊断性测量准确性的关键因素 （Cui et al., 2012）。一类错误率（type I error）也称“弃真”率，是指正常被试被误判为失拟被试的比率，统计检验力是指正确检测出失拟被试的比率。

实验设计：采用 2×2×6 三因素完全随机实验设计。三个因素分别为项目长度（20、40）、项目质量（高区分度、低区分度）、失拟被试类型（创造性作答、随机作答、疲劳、睡眠、作弊、随机作弊; Cui & Li, 2015; Santos et al., 2020）。其中高区分度项目设置成失误参数s和猜测参数g服从均匀分布 U(0.05,0.25)抽取，低区分度项目为失误参数s和猜测参数g服从均匀分布 U(0.25,0.40)抽取。根据 Cui 和 Li（2015），创造性作答指高能力的被试做错简单的项目，这里的高能力被试定义为掌握了所有考察属性的被试，简单的项目定义为只测量一个属性的项目，实验设置为假设每个被试掌握每个属性的概率为 80%，随机生成被试的属性掌握模式，被试在只测量一个属性的项目上答错；随机作答指测验动机低下的被试凭猜测随机作答，本研究设计为每个被试正确作答每题的概率为 25%（Yu & Cheng, 2019）；睡眠指考试中未能正确回答前几题，本研究设计为被试在前 25%题目上答错；疲劳指考试中未能正确回答后几题，这里设计为被试在后 25%题目上答错；作弊指低能力被试抄袭高能力被试的答案，从而答对较难项目，本研究设置为按 20%概率掌握各个属性的被试中，掌握 2 个

属性以下的被试在考察 3 个属性以上的项目上正确作答；随机作弊指低能力被试以 90%的概率答对 10%的难题（Santos et al., 2020）。

本研究控制变量包括：被试数量控制为 1000 人，选用认知诊断模型为常见的 DINA 模型，考察属性为 6 个，固定 Q 矩阵（Q 矩阵的详细信息请见附录 B）。被试知识状态和项目参数用 R 语言以 DINA 模型估计生成。重复实验 30 次，评价指标为一类错误率和统计检验力，检验水准$\alpha = 0.05$，本研究中一类错误率设置为不同实验条件下在 DINA 模型生成的 1000 个正常被试反应模式中，被指标误判为失拟被试的比例；统计检验力指标设置为每种异常被试类型生成 1000 个失拟被试，被鉴别出的异常被试的比例。取 30 次实验结果平均值作为最终评价指标。

l_z指标和RCI指标均在显著性水平为 0.05 的情况下，根据理论分布取临界值，l_z指标取 5 分位为临界值，RCI指标取 95 分位数为临界值。对于R指标，本研究采用经验临界值，具体做法是：给定 Q 矩阵，根据 DINA 模型，假设被试的知识状态服从均匀分布来估计被试知识状态，生成 10000 个正常被试作答数据，使用 MMLE/EM 估计项目参数（de la Torre, 2009），为每位被试计算R_i值，从低到高排序，取R_i值的 95 分位数作为临界值。

3.2 研究结果

表 1 给出了不同实验条件下三个指标的一类错误率和对不同异常被试类型的统计检验力，表 2 给出了不同测验长度下的模式判准率和属性边际判准率。一类错误率的结果显示，R指标对一类错误率控制得较好，稳定在 0.05，而l_z指标和RCI指标一类错误率出现了略微膨胀，在题目数量为 40 题时，RCI指标一类错误率趋于合理。这与 Cui 等（2015）研究结果中l_z指标和RCI指标一类错误率在正常范围有些不一致，原因可能是本研究采用的认知诊断模式为 DINA 模型，而 Cui 等人（2015）研究中使用的是 C-RUM 模型。

在统计检验力方面，随着题目区分度提高，各个指标在不同异常被试类型的统计检验力均有所提高，其中l_z指标在疲劳、睡眠、创造性作答和随机作答的异常被试类型下，随着题目区分度提高，统计检验力提升尤为明显，这个结果与 Cui 和 Li（2015）的结果一致。随着题目数量从 20 增加到 40，大部分统计检验力呈现上升趋势，但l_z指标在疲劳和睡眠的异常被试类型下，以及R指标在随机作弊的异常被试类型下，随着题目数量的增加，统计检验力有略微下降。

对于不同的异常被试类型，模拟研究结果显示在随机作答和随机作弊情况下，R指标表现最好，在疲劳、睡眠和创造性作答情况下l_z指标则表现更优，而随着题量增加，R指标在这三种情况下的统计检验力接近于l_z指标，这可以用随着题量的增加，模式判准率和属性判

准率都有所提高来解释。在低区分度题目上，在疲劳和睡眠的情况下，R指标比l_z指标和RCI指标表现更好。在作弊情况下，则是RCI指标表现最好且最稳定，l_z指标表现不理想。

表 1 不同异常被试类型的一类错误率和统计检验力

题目数量	题目区分度	指标	一类错误率	统计检验力					
				疲劳	睡眠	创造性作答	随机作答	作弊	随机作弊
20 题	高区分度	R	0.05 (0.01)	0.40 (0.01)	0.43 (0.02)	0.96 (0.01)	0.97 (0.01)	0.88 (0.01)	0.80 (0.02)
		l_z	0.08 (0.01)	0.85 (0.01)	0.81 (0.03)	1 (0)	0.62 (0.02)	0.11 (0)	0.66 (0.01)
		RCI	0.06 (0.01)	0.25 (0.01)	0.41 (0.05)	0.91 (0.01)	0.95 (0.02)	1 (0)	0.18 (0.01)
	低区分度	R	0.05 (0.01)	0.38 (0.02)	0.35 (0.01)	0.77 (0.03)	0.85 (0.01)	0.76 (0.02)	0.79 (0.01)
		l_z	0.08 (0.01)	0.09 (0)	0.03 (0.01)	0 (0)	0.29 (0.01)	0.01 (0)	0.01 (0)
		RCI	0.06 (0.01)	0.07 (0)	0.24 (0.03)	0.90 (0.02)	0.78 (0.03)	1 (0)	0.09 (0.01)
40 题	高区分度	R	0.05 (0.01)	0.63 (0.01)	0.72 (0.01)	1 (0)	1 (0)	0.95 (0.01)	0.77 (0.02)
		l_z	0.06 (0.01)	0.78 (0.01)	0.74 (0.02)	1 (0)	0.68 (0.01)	0.10 (0.01)	0.72 (0.01)
		RCI	0.05 (0.01)	0.43 (0.02)	0.66 (0.01)	1 (0)	0.99 (0.01)	1 (0)	0.20 (0.01)
	低区分度	R	0.05 (0.01)	0.51 (0.01)	0.34 (0.02)	0.87 (0.03)	0.84 (0.01)	0.78 (0.02)	0.50 (0.02)
		l_z	0.07 (0.01)	0.07 (0.01)	0.07 (0.01)	0.02 (0.01)	0.60 (0.01)	0.01 (0)	0.01 (0)
		RCI	0.05 (0.01)	0.08 (0.01)	0.25 (0.01)	0.89 (0.02)	0.79 (0.03)	1 (0)	0.10 (0.01)

备注：括号内数值表示重复实验 30 次的标准差。

综合可以看出：随着题目数量和题目质量的增加，各个被试拟合指标对异常被试类型侦察度越好，创造性作答的异常被试类型较容易被检测出来；RCI指标适合检测作弊的异常被试类型；l_z指标则更适合检测疲劳、睡眠的异常被试类型；R指标对创造性作答、随机作答和作弊均有较好的统计检验力，且在低区分度的题目上，R指标表现也最稳健。

表 2 模式判准率与属性判准率

题量	判准率	
	模式判准率（PCCR）	属性判准率（ACCR）
20 题	0.48	0.87
40 题	0.77	0.95

4. 研究二，*R*指标在实证数据中的应用研究

教育评估工具应该能反映学生的学习状态，为教学改进提供反馈信息。认知诊断评估对被试在测验所考察属性上的掌握水平进行分类，确定被试对哪些属性已经掌握，对哪些属性需要补救，而被试拟合检验能更好的确保被试评估分类的准确性和有效性。为了进一步检验*R*指标在认知诊断评估中应用的可行性，本部分将以分数减法的数据为例，用*R*指标进行被试拟合检验与分析。

4.1 实测数据来源

本研究采用实测数据为众多实证研究中运用的 Tatsuoka 分数减法数据，共 536 个被试，题目数量为 11 题（Henson et al., 2009）。该测验共考察 3 个属性，A1 从整数借位（borrowing from whole number）, A2 将整数和分数分开（separating whole number from fraction）, A3 通分（finding common denominator），其测验 Q 矩阵见表 3。

表 3 实证研究数据 Q 矩阵

题目编号	题目	属性		
		A1	A2	A3
1	3 1/2-2 3/2	1	1	0
2	3-2 1/5	1	0	1
3	3 7/8-2	1	0	1
4	4 4/12-2 7/12	1	0	0
5	4 1/3-2 4/3	1	1	0
6	11/8-1/8	1	1	0
8	2-1/3	1	0	1
9	4 5/7-1 4/7	1	0	1
10	7 3/5-4/5	1	0	0
11	4 1/10-2 8/10	1	0	0
13	4 1/3-1 5/3	1	1	0

4.2 研究过程与方法

本研究根据分数减法（Henson et al., 2009）的 Q 矩阵和作答数据，采用 DINA 模型，通过 R 语言中的 GDINA 包估计出项目参数和被试的属性掌握模式，项目参数结果见表 4。再根据估计出的项目参数，模拟 10000 名正常被试的作答数据，取 95 分位数为判别异常被试的临界值，最后根据R指标的临界值对实际作答数据进行被试拟合检验。并且，将RCI指标和l_z指标也应用到这批数据上，比较它们在分析结果上的差异。RCI和l_z指标的结果说明请见附录 C。

表 4 实证研究数据项目参数

题目编号	项目参数	
	失误参数s	猜测参数g
1	0.1207	0.2158
2	0.1762	0.1069
3	0.1513	0.5088
4	0.2487	0.0321
5	0.0691	0.0677
6	0.0465	0.5304
8	0.0864	0.1333
9	0.0534	0.5233
10	0.1462	0.0329
11	0.1577	0.1129
13	0.1762	0.0078

4.3 研究结果

结果显示，有 23 名被试被检测出作答反应异常，占总人数的 4.29%。下表列出了部分异常反应被试的基本情况。

编号为 24、48 和 97 号的被试答对了第 5、6、9、10 题，这几题考察第 1 个属性 4 次，考察第 2 个属性 2 次，未考察第 3 个属性，估计这几个被试属性掌握模式为[110]，其理想作答反应为[10011100111，但被试均在第 1、4、11 题上答错，第 1 题和第 11 题考察属性 A1 和 A2，可能还需进一步分析被试是否掌握了第 2 个属性。

编号为 137 号被试观察作答反应是[00001011111]，估计其属性掌握模式为[111]， 从理论上讲被试掌握了所有的属性，那么他在所有题目上都应该答对，但被试在实际上前四

题都答错了，有可能出现了“睡眠”的异常反应模式。

编号为 230 号被试，模型估计其属性掌握模式为[000]，但其观察作答反应为[01100100110]，答对了第 2、3、6、9、10 题，有可能有作弊行为。

表 5 部分异常反应被试情况

被试编号	观察作答	理想作答	属性掌握模式
24、48、97	00001100110	10011100111	110
25	10111100000	10011100111	110
37	10000111101	11111111111	111
63	01101101010	11111111111	111
115	11010001010	01110011110	101
137	00001011111	11111111111	111
171、194	00000101111	10011100111	110
183	11101001100	11111111111	111
203	01011111001	11111111111	111
219	01110110000	00000000000	000
230	01100100110	00000000000	000
449	10011101000	10011100111	110
...	...	...	...

5. 讨论与进一步的研究方向

本研究新提出认知诊断评估中的被试拟合指标R指标，并将其与l_z指标和RCI指标进行比较。在模拟研究中，R指标一类错误率稳定在 0.05 左右，较为正常，可用于认知诊断评价中对异常反应被试进行侦察。首先，模拟研究结果表明，随着题目数量增加和题目区分度提高，R指标检测率越高，这与预期一致。但本研究中，l_z指标一类错误率出现略微膨胀以及l_z指标在疲劳和睡眠两种异常被试类型下，随着题目数量的增加，统计检验力出现下降的现象，与 Cui 等人（2015）研究结果不一致，可能是所选模型不同导致，有待进一步研究加以探讨。

其次，由于目前还不完全了解R指标的理论分布，本研究中R指标的临界值是采用经验分布确定的，这在实际应用中可能不方便使用，探索R指标的统计性质，如果能够推导出它的理论零分布或近似分布（Andrews, 1993），则更有助于它的应用和推广。

第三，本研究中的R指标是对各考生所有项目上的求和，如果将R指标定义在各项目在所有考生上的求和，则可以用于项目拟合检验（Drasgow et al., 1985），因此，将R指标推广到项目拟合检验也是值得研究和探索的。

第四，项目质量对于被试拟合检验有非常大的影响，本研究没有把项目质量纳入考虑是一个不足之处，未来需进一步探索项目质量对于R指标的表现。除此之外，Wang 等（2018）对确定被试异常作答的类型进行了尝试，这方面的工作也需要进行深入的探索。在实证研究中，由于采用的是其他研究的实证数据，故无法对侦察出的异常被试做进一步分析以及补救措施。而且被试出现异常作答反应的原因不能仅仅只根据被试拟合指标来确定，因为被试拟合指标不能直接指出异常反应行为的实际原因，因此，进一步分析被试考试行为的辅助信息如被试的口头报告、座位安排、考试时间等是十分必要的。

最后，由于二级计分方式的模型只能评价被试是否掌握某一知识或技能，而对被试在不同知识或技能的掌握水平或程度不能进行有效地评价，在实际情景中，教育与心理测验中的题目形式丰富多样，如教育考试中的计算题、论述题、简答题、证明题、作文题，心理量表中的 Likert 型量表等等，这些题型的数据基本都是多级评分数据（丁树良等, 2014; 夏梦连等, 2018; 王鹏等, 2019）或多分属性下的评分数据（丁树良等, 2015; 詹沛达等, 2017），未来可将被试拟合检验扩展到多级计分或多分属性下的认知诊断。

6 结论

本研究提出在认知诊断框架下的被试拟合指标R，通过模拟研究比较RCI、l_z和R指标的一类错误率及统计检验力，并将指标应用于实证数据，验证R指标在实证数据中的表现。研究结果表明，R指标一类错误率较合理，l_z指标和RCI指标一类错误率出现了略微膨胀。随着题目区分度和题目数量的增加，指标的统计检验力有所提高。对于不同的异常被试类型，RCI指标适合检测作弊的异常被试类型，l_z指标适合检测疲劳、睡眠的异常被试类型，R指标对创造性作答、随机作答和作弊的异常被试均有较好的侦察力。

参考文献

陈孚， 辛涛， 刘彦楼， 刘拓， 田伟． (2016)． 认知诊断模型资料拟合检验方法和统计量. *心理科学进展, 24*(12)， 1946–1960.

丁树良， 汪文义， 罗芬． (2014)． 多级评分认知诊断测验蓝图的设计——根树型结构． *江西师范大学学报(自然科学版), 38*(2)， 111-118.

丁树良， 汪文义， 罗芬， 熊建华． (2015)． 多值 Q矩阵理论． *江西师范大学学报(自然科学版), 39*(4)， 365-370.

涂冬波，张心，蔡艳，戴海琦．(2014)．认知诊断模型-资料拟合检验统计量及其性能．*心理科学，37*(1)，205–211.

王鹏,孟维璇,朱干成,张登浩,张利会,董一萱,司英栋.(2019).多维项目反应理论补偿性模型参数估计:基于广义回归神经网络集合．*心理学探新, 39*(3),244-249.

夏梦连，毛秀珍，杨睿．(2018)．属性多级和项目多级评分的认知诊断模型．*江西师范大学学报(自然科学版)，42*(2)，134-138.

詹沛达，丁树良，王立君．(2017)．多分属性层级结构下引入逻辑约束的理想掌握模式．*江西师范大学学报(自然科学版)，41*(3)，289-295.

Andrews, D. W. K. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica, 61*(4), 821–856.

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). Chapman & Hall/CRC.

Cui, Y., & Li, J. (2015). Evaluating person fit for cognitive diagnostic assessment. *Applied Psychological Measurement, 39*(3), 223-238.

Cui, Y., Gierl, M. J., & Chang, H.-H. (2012). Estimating classification consistency and accuracy for cognitive diagnostic assessment. *Journal of Educational Measurement, 49*(1), 19–38.

Cui, Y., & Leighton, J. P. (2009). The hierarchy consistency index: Evaluating person fit for cognitive diagnostic assessment. *Journal of Educational Measurement, 46*(4), 429–449.

Cui, Y., & Li, J. (2015). Evaluating person fit for cognitive diagnostic assessment. *Applied Psychological Measurement, 39*(3), 223–238.

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics, 34*(1), 115–130.

Drasgow, F, Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*(1), 67–86.

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*(2), 191–210.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*(3), 258–272.

Leighton, J. P., & Gierl, M. J. (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge University Press.

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational and Behavioral Statistics, 4*(4), 269–290.

Liu, Y., Douglas, J. A., & Henson, R. A. (2009). Testing person fit in cognitive diagnosis. *Applied Psychological Measurement, 33*(8), 579–598.

Masters, G. N., & Wright, B. D. (1997). The partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 101–121). Springer.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25*(2), 107–135.

Molenaar, I. W., & Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika, 55*(1), 75–106.

Reise, S. P. (1995). Scoring method and the detection of person misfit in a personality assessment context. *Applied Psychological Measurement, 19*(3), 213–229.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications.* Guilford Press.

Santos, K. C. P., de la Torre, J., & von Davier, M. (2020). Adjusting person fit index for skewness in cognitive diagnosis modeling. *Journal of Classification, 37*(2), 399–420.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*(4), 345–354.

von Davier, M., & Lee, Y.-S. (2019). *Handbook of diagnostic classification models*. Cham: Springer International Publishing.

Wang, C., Xu, G. J., & Shang, Z. R. (2018). A two-stage approach to differentiating normal and aberrant behavior in computer based testing. *Psychometrika, 83*(1), 223–254.

Yu, X. F., & Cheng, Y. (2019). A change-point analysis procedure based on weighted residuals to detect back random responding. *Psychological Methods, 24*(5), 658–674.

Research on Person-fit in Cognitive Diagnostic Assessment

Yu Xiaofeng[1]*, Tang Qian*[1]*, Qin Chunying*[2]*, Li Yujun*[1]

([1]School of Psychology, Jiangxi Normal University, Nanchang, 330022)

([2]School of Mathematics and Information Science，Nanchang Normal University，Nanchang 330032)

Abstract Cognitive Diagnostic Assessment (CDA) has been widely used in educational assessment. It can provide guidance for further study and teaching by analyzing whether the test-takers have acquired knowledge points or skills.

In psychometrics, statistical methods for assessing the fit of an examinee's item responses to a postulated psychometric model are often called person-fit statistic. The person-fit analysis can help to verify the individual diagnostic results, and is mainly used to distinguish the abnormal examinees from the normal ones. The abnormal response patterns include "sleeping" behavior, fatigue, cheating, creative responding, random guessing responses and cheating with randomness, and all of these abnormal response patterns can affect the deviation of examinee's ability estimation. The person-fit analysis can help researchers identify the abnormal response patterns more accurately, so as to delete the abnormal responding examinees and improve the validity of the test. In the past, most of the person fit researches were mainly carried out under the Item Response Theory (IRT) framework, while only few papers have been published dealing with person-fit under the CDM framework. This study attempts to fill a gap in the literature by introducing new methods. In this study, a new person fit index (R) was proposed.

In order to verify the validity of the newly developed person fit index, this study explores the type I error and statistical test power of R index under different item length, item discrimination and different misfit types of respondent, and compares it with existing methods RCI and l_z. Type I error rate was defined as the proportion of flagged abnormal response patterns by a person fit statistic out of 1,000 generated normal response patterns from the DINA model. The control variables of this study include: the number of subjects is controlled to 1000, the cognitive diagnosis model is chosen as DINA model, the attributes are 6, and the Q matrix is fixed. Finally, in order to reflect the value of person fit index in practical application, the R index is applied to the empirical data of fractional subtraction.

The results show that the type I error of R index is reasonable and stable at 0.05. In the aspect of statistical test power, with the improvement of item differentiation, the statistical test power of each index in different abnormal examinees is improved. With the increase in the number of items, most of the statistical power show an upward trend. For different types of abnormal subjects, R index perform best in the cases of random guessing responses and cheating with randomness. In the case of fatigue, sleep, and creative responding, the l_z index perform better. In the empirical data study, the detection rate of abnormal examinees is 4.29%.

With the increase of the discrimination of items and the increase of the number of items, the power of R index has improved, and the performance of R index is the most robust when the discrimination of item is low. The R index has a high power for the types of abnormal behavior such as creative responding behavior, random guessing responses and cheating with randomness.
Keywords cognitive diagnosis, person fit, DINA model, aberrant response

附录 A　l_z和RCI指标

(1) l_z指标

Cui 和 Li（2015）将l_z指数（Drasgow et al., 1985）引入到认知诊断测验中，l_z指标是常见的项目反应理论下的被试拟合指标，源于似然函数l_0（Levine & Rubin, 1979），是l_0的标准化。基于选定的项目反应理论（IRT）模型，l_0计算观察的项目反应模式的对数似然值，表达式如下：

$$l_{0i} = ln\{\prod_{j=1}^{J} P_j(\theta_i)^{X_{ij}}[1 - P_j(\theta_i)]^{1-X_{ij}}\} \quad \text{(A-1)}$$

其中X_{ij} 是两级（0, 1）计分，表示被试i在第$j(j = 1,2,\dots,J)$个项目的观察反应，$P_j(\theta_i)$ 是能力为θ_i的被试i在项目j上的正确作答概率，l_{0i}较小值表示给定的 IRT 模型中，能力为θ_i的被试i出现反应模式X_i的概率较小。对l_{0i}进行标准化，得到统计量l_z为：

$$l_z = \frac{l_0 - E(l_0)}{[Var(l_0)]^{1/2}} \quad \text{(A-2)}$$

其中，$E(l_0) = \sum_{j=1}^{J}\{P_j(\theta)\, ln[P_j(\theta)] + [1 - P_j(\theta)]\, ln[1 - P_j(\theta)]\}$　(A-3)

$$Var(l_0) = \sum_{j=1}^{J} P_j(\theta)[1 - P_j(\theta)]\left[ln\frac{P_j(\theta)}{1-P_j(\theta)}\right]^2 \quad \text{(A-4)}$$

Cui 和 Li（2015）将$P_j(\theta_i)$ 改为认知诊断模型中的$P_j(\alpha_i)$，构建了基于诊断测验中的拟合统计量l_z，其模拟研究中发现，基于被试属性掌握模式估计值时，l_z指标呈负偏态分布，这与l_z指标在项目反应理论下的结果一致（Molenaar & Hoijtink, 1990; Reise, 1995）。

(2) 反应一致性指标RCI

由于HCI指标是依赖于属性之间的层级关系的，当属性之间不存在属性层级关系时，HCI指标就不能使用。基于此，Cui 和 Li（2015）提出了反应一致性指标（response conformity index, RCI），RCI是评估 Q 矩阵的预测反应和被试观察反应之间的一致性程度的，其表达式如下：

$$RCI_i = \sum_{j=1}^{J}|RCI_{ij}| = \sum_{j=1}^{J}\left|ln\left[-\frac{X_{ij}-P_j(\alpha_i)}{I_j(\alpha_i)-P_j(\alpha_i)}\right]^{X_{ij}+I_j(\alpha_i)}\right| \quad \text{(A-5)}$$

其中，α_i为被试i的属性掌握模式；$P_j(\alpha_i)$ 为属性掌握模式α_i的被试正确作答项目j的概率；$I_j(\alpha_i)$为属性掌握模式α_i的被试对项目j的理想反应。即当被试掌握了项目考察的所有属性时，$I_j(\alpha_i) = 1$，如果缺少一个或多个属性，则该项目的理想反应将为 0；X_{ij}为被试的实际作答，取值为 0 或 1。

对于每个题目，RCI_i测量的是观察到的项目反应X_{ij}与理想反应$I_j(\alpha_i)$ 的偏离程度，当$X_{ij} = I_j(\alpha_i)$时，$RCI_i = 0$，说明被试拟合很好；当$X_{ij} \neq I_j(\alpha_i)$ 时，被试拟合取决于$X_{ij} - P_j(\alpha_i)$

和$I_j(\alpha_i) - P_j(\alpha_i)$ 的差异大小，如果$X_{ij} - P_j(\alpha_i)$ 和$I_j(\alpha_i) - P_j(\alpha_i)$相比比较大，表明被试对项目的作答是不符合期望的，因此，可能出现了异常反应行为，如作弊、创造性作答，此时RCI为一个较大的正值，相反，如果$I_j(\alpha_i) - P_j(\alpha_i)$与$X_{ij} - P_j(\alpha_i)$相比比较大，可能的原因是题目质量较差，或被试作答时采用了非 Q 矩阵指定的策略，这种情况下，RCI为一个较大的负值。

附录 B　研究一所使用的 Q 矩阵

表 B-1 $K=6, J=20$模拟数据的 Q 矩阵

题目	α_1	α_2	α_3	α_4	α_5	α_6	题目	α_1	α_2	α_3	α_4	α_5	α_6
1	1	0	0	0	0	0	11	0	1	0	1	1	0
2	0	1	0	0	0	0	12	1	1	1	1	1	0
3	0	0	1	0	0	0	13	1	1	1	1	1	0
4	0	0	0	1	0	0	14	1	0	1	0	1	0
5	0	0	0	0	1	0	15	0	0	1	1	0	1
6	0	0	0	0	0	1	16	0	0	1	0	1	0
7	0	1	0	0	1	1	17	1	0	0	1	1	1
8	0	0	0	0	1	1	18	0	1	1	0	0	0
9	1	1	1	1	0	0	19	1	0	1	0	0	1
10	0	0	1	0	1	1	20	0	1	0	1	1	0

表 B-2 $K=6, J=40$模拟数据的 Q 矩阵

题目	α_1	α_2	α_3	α_4	α_5	α_6	题目	α_1	α_2	α_3	α_4	α_5	α_6
1	1	0	0	0	0	0	21	1	0	0	1	1	0
2	0	1	0	0	0	0	22	1	1	1	0	1	1
3	0	0	1	0	0	0	23	0	0	0	0	1	1
4	0	0	0	1	0	0	24	0	0	1	0	0	1
5	0	0	0	0	1	0	25	0	0	1	0	1	1
6	0	0	0	0	0	1	26	0	1	0	1	1	0
7	1	0	1	1	1	0	27	0	0	1	0	1	0
8	0	1	0	0	0	0	28	0	0	0	0	0	1
9	0	1	1	0	1	0	29	0	0	0	1	0	1
10	0	0	0	0	1	0	30	1	0	1	1	0	1
11	1	1	1	1	0	0	31	0	0	0	1	1	1
12	1	0	0	1	0	1	32	0	0	0	0	1	0
13	1	0	0	0	0	1	33	0	1	0	1	1	1
14	0	1	1	1	0	1	34	0	0	0	1	1	0
15	1	0	0	1	0	1	35	1	0	1	0	1	0
16	0	1	1	0	1	1	36	1	0	0	0	1	1
17	1	0	0	1	0	1	37	0	1	0	0	0	1
18	0	0	0	0	1	0	38	1	0	0	0	0	1
19	1	0	0	1	0	0	39	1	1	1	0	0	1
20	0	0	1	1	1	0	40	1	0	0	0	1	0

附录 C RCI和l_z指标对分数减法数据的分析结果说明

除了R指标外，同时也将RCI和l_z指标对分数减法数据进行了分析。结果显示，l_z和RCI指标分别检测出 47 和 35 名被试反应异常，占总人数的 8.8%和 6.5%。这里面比较有趣的现象是R指标虽然只标记出了 23 位异常考生，但是这 23 人里面有 1 人没有被l_z标记出来，全部被RCI标记出来，这个结果也表明R指标在标记异常考生的时候会更“保守”，这对于“高风险”的测验来说是非常有必要的，因为在标记考生异常作答行为时需要非常慎重，通常要综合多种方法来对考生的数据进行分析，然后才能做出决策。进一步，没有被l_z标记出但是被l_z标记出的考生编号为 137，其观察得分模式[00001011111]，可以看出，这位考生出现了类似“热身效应”或“睡眠”的作答行为，这个结果也从一个侧面验证了l_z指标在较短的测验中对于睡眠行为的检验力严重依赖题目质量的特点。

比较RCI和l_z指标的分析结果，被它们同时标出的考生有 28 人，分别占各自标出总人数的 80%和 60%。这个比例也进一步说明l_z指标在标记异常考生时标准相对宽松，与研究一中l_z略微“膨胀”的一类错误率是对应的。